## Supplementary Methods on Multiple Imputation

### Sensitivity analysis (text reproduced from body of main paper)

To address the impact of missing covariate data (5.8% of individuals missing ethnicity and/or NZDep quintile), we used multiple imputation to examine whether the associations measured in the main analysis could have been biased due to exclusion of individuals with missing data (complete case analysis). Five imputation datasets were created using chained equations[32] (using the mice package in R[33]). These datasets imputed missing values for ethnicity and NZDep quintile (as polynomial variables) based on all other variables in the analytical model including exposure variables and outcome variables (multimorbidity status, age group, sex, ethnicity, NZDep quintile, and all outcome variables). The imputation models also included auxiliary information on each person's District Health Board of residence (the 20 administrative divisions of the public health system in NZ, which provides additional information on sub-national distribution of people by ethnicity and socioeconomic deprivation). Further details on this analysis and underlying assumptions are given with Supplementary Table B.

### References from main paper:


32. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011;30(4):377-99. doi: 10.1002/sim.4067 [published Online First: 2011/01/13]
33. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of statistical software* 2011;45(3):1-67.


## Supplementary Methods on Assumptions of Multiple Imputation

The following notes assume some familiarity with methods for missing data and multiple imputation: several overview papers have been previously published on this methodology[1-3].

In order for multiple imputation of covariates to be valid and useful, a key assumption is that data are missing at random (MAR), which means that the to-be-imputed values can be considered to be missing at random conditional on the variables included in the imputation model. [1 2] Thus, an imputation process that draws on these conditioning variables (including exposure and outcome variables) to produce imputed values should be able to recover some information to account for the potential profile of those people who are missing some data. It is not possible to determine from a dataset whether data are missing at random or missing not at random (MNAR: i.e. some additional unmeasured information influences whether data are missing). [2 3] However, including a sufficient number of meaningful variables as predictors in the imputation model process, including exposure and outcome variables, serves to make the missing at random assumption more plausible for a given scenario[1 3].

In the current study, we believe on theoretical grounds that the missing data (for ethnicity and socioeconomic status as measured by area of residence using NZDep 2013) are effectively missing at random, conditional on the variables included in our imputation model.

Firstly, we assume that ethnicity data collected in the routine data sources is more likely to be present for people with multiple health contacts (because these are opportunities to collect ethnicity data in line with NZ's ethnicity data protocols). The imputation models explicitly include information on multimorbidity status and subsequent health outcomes in the imputation process. This means health-status is being used as part of the imputation process, which should lead to valid results for the imputation analysis (in conjunction with other known sources of patterning for ethnicity across NZ, including geographic variation and variation of socioeconomic status by ethnicity).

Secondly, NZDep values (the second missing variable in the regression models) tend to be missing when address information for a given person is either unavailable or incompletely recorded in the Ministry of Health's master databases (and hence geocoding cannot be performed to assign that person with an area-based code), or when there an otherwise-correct address cannot be mapped to the area codes recorded in the measure NZDep. The chances of this second scenario depend upon the discrepancy between the time at which a person's address is measured (usually the most recent update to their health record) and the timing of the specific five-yearly census from which the NZDep measure was derived (in this case, the 2013 census conducted in March 2013).

Supplementary Table B below includes both the complete-cases results of the regression models (top half, reproducing results from Table 4 of the main paper) and also the results of the analysis of the multiply-imputed datasets (bottom half of Sup. Table B) following the analytical procedures given in the main paper (as reproduced above). As can be seen, and as reported in the main paper, the results are almost identical in the two analyses: point estimates are marginally higher in the imputed-data results, but not substantively different.

**References for Supplementary Methods text:**


1. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59(10):1087-91. doi: 10.1016/j.jclinepi.2006.01.014 [published Online First: 2006/09/19]
2. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393. doi: 10.1136/bmj.b2393 [published Online First: 2009/07/01]
3. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. *Stat Med* 2007;26(16):3057-77. doi: 10.1002/sim.2787

Supplementary Table B. Results from original complete-case analysis (top panel, Table 4 from main paper) and from analysis of multiply imputed data (n=5 imputation datasets).

| | Odds ratio (95% CI) for risk of outcome with multimorbidity* | | | | | |
|---|---|---|---|---|---|---|
| | Hospital discharge definition | | | Pharmaceutical dispensing definition | | |
| Model† | Mortality | ASH‡ | Admission§ | Mortality | ASH‡ | Admission§ |
| COMPLETE CASE ANALYSIS | | | | | | |
| Unadjusted model | 17.6 (17.2, 18.1) | 8.4 (8.3, 8.5) | 5.6 (5.6, 5.7) | 14.7 (14.2, 15.2) | 5.5 (5.5, 5.6) | 3.7 (3.7, 3.7) |
| Adjusted age, sex | 4.8 (4.7, 5.0) | 4.9 (4.9, 5.0) | 3.6 (3.5, 3.6) | 4.0 (3.9, 4.2) | 3.6 (3.6, 3.7) | 2.6 (2.6, 2.7) |
| + adjust ethnicity | 4.7 (4.6, 4.8) | 4.7 (4.6, 4.7) | 3.5 (3.5, 3.5) | 3.9 (3.8, 4.1) | 3.6 (3.5, 3.6) | 2.6 (2.6, 2.6) |
| + adjust NZDep quintile | 4.6 (4.5, 4.7) | 4.6 (4.5, 4.6) | 3.5 (3.4, 3.5) | 3.9 (3.7, 4.0) | 3.5 (3.5, 3.6) | 2.6 (2.6, 2.6) |
| MULTIPLE IMPUTATION ANALYSIS | | | | | | |
| Unadjusted model | 18.0 (17.5, 18.4) | 8.7 (8.6, 8.9) | 5.8 (5.8, 5.9) | 14.8 (14.3, 15.3) | 5.7 (5.6, 5.8) | 3.8 (3.8, 3.8) |
| Adjusted age, sex | 4.9 (4.8, 5.0) | 5.1 (5.1, 5.2) | 3.7 (3.7, 3.7) | 4.1 (4.0, 4.2) | 3.7 (3.7, 3.8) | 2.7 (2.7, 2.7) |
| + adjust ethnicity | 4.8 (4.6, 4.9) | 4.8 (4.8, 4.9) | 3.6 (3.6, 3.7) | 4.0 (3.9, 4.1) | 3.7 (3.6, 3.7) | 2.7 (2.7, 2.7) |
| + adjust NZDep quintile | 4.7 (4.6, 4.8) | 4.7 (4.7, 4.8) | 3.6 (3.6, 3.6) | 3.9 (3.8, 4.1) | 3.6 (3.6, 3.7) | 2.7 (2.6, 2.7) |

* Reference group is individuals without multimorbidity (i.e. either zero or only one long-term conditions identified)

† All models run on complete-case data only (n=3,288,646; total of n=201,101 missing ethnicity &/or NZDep)

‡ Ambulatory sensitive hospitalisation (ASH)

§ Non-maternity admissions with at least an overnight stay.

Note: Complete-cases analysis reproduces results shown in Table 4 of main paper (regression results for people with complete data for all covariates included in the fully-adjusted model). 5.8% of individuals were missing ethnicity and/or NZDep quintile data in the complete-case analysis.